DOCUMENT RESUME

ED 378 559 CS 011 979

AUTHOR Schafer, William D.; And Others
TITLE Test Quality for Use in Curricular and Instructional Decision Making in Reading. Reading Research Report No. 28.
INSTITUTION National Reading Research Center, Athens, GA.; National Reading Research Center, College Park, MD.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE 94
CONTRACT 117A20007
NOTE 29p.; For a related document, see CS 011 978.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Criterion Referenced Tests; Educational Change; Elementary Secondary Education; *Evaluation Methods; Instructional Innovation; *Literacy; *Reading Achievement; Reading Instruction; *Reading Tests; State Standards; *Student Evaluation; Test Construction; *Test Use; Test Validity
IDENTIFIERS Maryland School Performance Assessment Program; *Performance Based Evaluation

ABSTRACT

Three expert panels reviewed the 1991 Maryland School Performance Assessment Program (MSPAP) reading test. This was the first year of an assessment program designed to measure school progress toward, among other content areas, three reading outcomes: reading for literary experience, reading to become informed, and reading to perform a task. The MSPAP, given throughout the state, is a nontraditional, criterion-referenced performance assessment, which in 1991 required 9 hours of testing time over an 8-day period. The three panels, one consisting of experts with an instructional perspective, one with a curricular perspective, and one a psychometric perspective, independently addressed a variety of test quality issues after review of the test materials and a presentation by a test developer. Results indicate that a test such as the 1991 MSPAP is adequate to assess school progress in reading but may be confounded with writing and may not adequately measure progress in basic reading skills. Findings suggest that a test such as the MSPAP is useful for making curricular and instructional decisions, but that use of the test for making decisions about individual students was not supported. (Contains 24 references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 378 559

Test Quality for Use in Curricular and Instructional Decision Making in Reading

William D. Schafer
John T. Guthrie
University of Maryland College Park

Janice F. Almasi
State University of New York at Buffalo

Peter P. Afflerbach
University of Maryland College Park

CS011979

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

NRRC
National Reading Research Center

READING RESEARCH REPORT NO. 28
Fall 1994

BEST COPY AVAILABLE

NRRC

National Reading Research Center

Test Quality for Use in Curricular and Instructional Decision Making in Reading

William D. Schafer
John T. Guthrie
University of Maryland College Park

Janice F. Almasi
State University of New York at Buffalo

Peter P. Afflerbach
University of Maryland College Park

READING RESEARCH REPORT NO. 28
Fall 1994

The work reported herein is a National Reading Research Project of the University of Georgia and University of Maryland. It was supported under the Educational Research and Development Centers Program (PR/AWARD NO. 117A20007) as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The findings and opinions expressed here do not necessarily reflect the position or policies of the National Reading Research Center, the Office of Educational Research and Improvement, or the U.S. Department of Education.

NRRC National Reading Research Center

Executive Committee
Donna E. Alvermann, Co-Director
University of Georgia
John T. Guthrie, Co-Director
University of Maryland College Park
James F. Baumann, Associate Director
University of Georgia
Patricia S. Koskinen, Associate Director
University of Maryland College Park
Nancy B. Mizelle, Acting Associate Director
University of Georgia
Jamie Lynn Metsala, Interim Associate Director
University of Maryland College Park
Penny Oldfather
University of Georgia
John F. O'Flahavan
University of Maryland College Park
James V. Hoffman
University of Texas at Austin
Cynthia R. Hynd
University of Georgia
Robert Serpell
University of Maryland Baltimore County
Betty Shockley
Clarke County School District, Athens, Georgia
Linda DeGroff
University of Georgia

Publications Editors

Research Reports and Perspectives
Linda DeGroff, Editor
University of Georgia
James V. Hoffman, Associate Editor
University of Texas at Austin
Mariam Jean Dreher, Associate Editor
University of Maryland College Park
Instructional Resources
Lee Galda, *University of Georgia*
Research Highlights
William G. Holliday
University of Maryland College Park
Policy Briefs
James V. Hoffman
University of Texas at Austin
Videos
Shawn M. Glynn, *University of Georgia*

NRRC Staff
Barbara F. Howard, Office Manager
Kathy B. Davis, Senior Secretary
University of Georgia

Barbara A. Neitzey, Administrative Assistant
Valerie Tyra, Accountant
University of Maryland College Park

National Advisory Board
Phyllis W. Aldrich
Saratoga Warren Board of Cooperative Educational Services, Saratoga Springs, New York
Arthur N. Applebee
State University of New York, Albany
Ronald S. Brandt
Association for Supervision and Curriculum Development
Marshá T. DeLain
Delaware Department of Public Instruction
Carl A. Grant
University of Wisconsin-Madison
Walter Kintsch
University of Colorado at Boulder
Robert L. Linn
University of Colorado at Boulder
Luis C. Moll
University of Arizona
Carol M. Santa
School District No. 5
Kalispell, Montana
Anne P. Sweet
Office of Educational Research and Improvement, U.S. Department of Education
Louise Cherry Wilkinson
Rutgers University

Production Editor
Katherine P. Hutchison
University of Georgia

Dissemination Coordinator
Jordana E. Rich
University of Georgia

Text Formatter
Ann Marie Vanstone
University of Georgia

NRRC - University of Georgia
318 Aderhold
University of Georgia
Athens, Georgia 30602-7125
(706) 542-3674 Fax: (706) 542-3678
INTERNET: NRRC@uga.cc.uga.edu

NRRC - University of Maryland College Park
2102 J. M. Patterson Building
University of Maryland
College Park, Maryland 20742
(301) 405-8035 Fax: (301) 314-9625
INTERNET: NRRC@umail.umd.edu

About the National Reading Research Center

The National Reading Research Center (NRRC) is funded by the Office of Educational Research and Improvement of the U.S. Department of Education to conduct research on reading and reading instruction. The NRRC is operated by a consortium of the University of Georgia and the University of Maryland College Park in collaboration with researchers at several institutions nationwide.

The NRRC's mission is to discover and document those conditions in homes, schools, and communities that encourage children to become skilled, enthusiastic, lifelong readers. NRRC researchers are committed to advancing the development of instructional programs sensitive to the cognitive, sociocultural, and motivational factors that affect children's success in reading. NRRC researchers from a variety of disciplines conduct studies with teachers and students from widely diverse cultural and socioeconomic backgrounds in pre-kindergarten through grade 12 classrooms. Research projects deal with the influence of family and family-school interactions on the development of literacy; the interaction of sociocultural factors and motivation to read; the impact of literature-based reading programs on reading achievement; the effects of reading strategies instruction on comprehension and critical thinking in literature, science, and history; the influence of innovative group participation structures on motivation and learning; the potential of computer technology to enhance literacy; and the development of methods and standards for alternative literacy assessments.

The NRRC is further committed to the participation of teachers as full partners in its research. A better understanding of how teachers view the development of literacy, how they use knowledge from research, and how they approach change in the classroom is crucial to improving instruction. To further this understanding, the NRRC conducts school-based research in which teachers explore their own philosophical and pedagogical orientations and trace their professional growth.

Dissemination is an important feature of NRRC activities. Information on NRRC research appears in several formats. *Research Reports* communicate the results of original research or synthesize the findings of several lines of inquiry. They are written primarily for researchers studying various areas of reading and reading instruction. The *Perspective Series* presents a wide range of publications, from calls for research and commentary on research and practice to first-person accounts of experiences in schools. *Instructional Resources* include curriculum materials, instructional guides, and materials for professional growth, designed primarily for teachers.

For more information about the NRRC's research projects and other activities, or to have your name added to the mailing list, please contact:

Donna E. Alvermann, Co-Director
National Reading Research Center
318 Aderhold Hall
University of Georgia
Athens, GA 30602-7125
(706) 542-3674

John T. Guthrie, Co-Director
National Reading Research Center
2102 J. M. Patterson Building
University of Maryland
College Park, MD 20742
(301) 405-8035

NRRC Editorial Review Board

Patricia Adkins
University of Georgia

Peter Afflerbach
University of Maryland College Park

JoBeth Allen
University of Georgia

Patty Anders
University of Arizona

Tom Anderson
University of Illinois at Urbana-Champaign

Harriette Arrington
University of Kentucky

Irene Blum
Pine Springs Elementary School
Falls Church, Virginia

John Borkowski
Notre Dame University

Cynthia Bowen
Baltimore County Public Schools
Towson, Maryland

Martha Carr
University of Georgia

Suzanne Clewell
Montgomery County Public Schools
Rockville, Maryland

Joan Coley
Western Maryland College

Michelle Commeyras
University of Georgia

Linda Cooper
Shaker Heights City Schools
Shaker Heights, Ohio

Karen Costello
Connecticut Department of Education
Hartford, Connecticut

Karin Dahl
Ohio State University

Lynne Diaz-Rico
California State University-San Bernardino

Pamela Dunston
Clemson University

Jim Flood
San Diego State University

Dana Fox
University of Arizona

Linda Gambrell
University of Maryland College Park

Valerie Garfield
Chattahoochee Elementary School
Cumming, Georgia

Sherrie Gibney-Sherman
Athens-Clarke County Schools
Athens, Georgia

Rachel Grant
University of Maryland College Park

Barbara Guzzetti
Arizona State University

Jane Haugh
Center for Developing Learning Potentials
Silver Spring, Maryland

Beth Ann Herrmann
Northern Arizona University

Kathleen Heubach
University of Georgia

Susan Hill
University of Maryland College Park

Sally Hudson-Ross
University of Georgia

Cynthia Hynd
University of Georgia

Robert Jimenez
University of Oregon

Karen Johnson
Pennsylvania State University

James King
University of South Florida

Sandra Kimbrell
West Hall Middle School
Oakwood, Georgia

Kate Kirby
Gwinnett County Public Schools
Lawrenceville, Georgia

Sophie Kowzun
Prince George's County Schools
Landover, Maryland

Linda Labbo
University of Georgia

Rosary Lalik
Virginia Polytechnic Institute

Michael Law
University of Georgia

Sarah McCarthey
University of Texas at Austin

Veda McClain
University of Georgia

Lisa McFalls
University of Georgia

Mike McKenna
Georgia Southern University

Donna Mealey
Louisiana State University

Barbara Michalove
Fowler Drive Elementary School
Athens, Georgia

Akintunde Morakinyo
University of Maryland College Park

Lesley Morrow
Rutgers University

Bruce Murray
University of Georgia

Susan Neuman
Temple University

Caroline Noyes
University of Georgia

John O'Flahavan
University of Maryland College Park

Penny Oldfather
University of Georgia

Joan Pagnucco
University of Georgia

Barbara Palmer
Mount Saint Mary's College

Mike Pickle
Georgia Southern University

Jessie Pollack
Maryland Department of Education
Baltimore, Maryland

Sally Porter
Blair High School
Silver Spring, Maryland

Michael Pressley
State University of New York at Albany

Tom Reeves
University of Georgia

Lenore Ringler
New York University

Mary Roe
University of Delaware

Nadeen T. Ruiz
California State University-Sacramento

Rebecca Sammons
University of Maryland College Park

Paula Schwanenflugel
University of Georgia

Robert Serpell
University of Maryland Baltimore County

Betty Shockley
Fowler Drive Elementary School
Athens, Georgia

Susan Sonnenschein
University of Maryland Baltimore County

Steve Stahl
University of Georgia

Anne Sweet
Office of Educational Research and Improvement

Liqing Tao
University of Georgia

Ruby Thompson
Clark Atlanta University

Louise Tomlinson
University of Georgia

Sandy Tumarkin
Strawberry Knolls Elementary School
Gaithersburg, Maryland

Sheila Valencia
University of Washington

Bruce VanSledright
University of Maryland College Park

Chris Walton
Northern Territory University
Australia

Janet Watkins
University of Georgia

Louise Waynant
Prince George's County Schools
Upper Marlboro, Maryland

Priscilla Waynant
Rolling Terrace Elementary School
Takoma Park, Maryland

Dera Weaver
Athens Academy
Athens, Georgia

Jane West
Agnes Scott

Steve White
University of Georgia

Allen Wigfield
University of Maryland College Park

Shelley Wong
University of Maryland College Park

About the Authors

William D. Schafer is Associate Professor, Department of Measurement, Statistics, and Evaluation, College of Education, University of Maryland College Park. He specializes in applied assessment.

John T. Guthrie is Professor, Department of Human Development, College of Education, University of Maryland College Park. He is Co-Director of the National Reading Research Center.

Janice F. Almasi is Assistant Professor, Department of Learning and Instruction, State University of New York at Buffalo. She specializes in social construction of literacy and its effect on cognitive processing and engagement.

Peter P. Afflerbach is Associate Professor, Department of Curriculum and Instruction, College of Education, University of Maryland College Park. He specializes in reading assessment.

National Reading Research Center
Universities of Georgia and Maryland
Reading Research Report No. 28
Fall 1994

Test Quality for Use in Curricular and Instructional Decision Making in Reading

William D. Schafer
John T. Guthrie
University of Maryland College Park

Janice F. Almasi
State University of New York at Buffalo

Peter P. Afflerbach
University of Maryland College Park

Abstract. *Three expert panels reviewed the 1991 Maryland School Performance Assessment Program (MSPAP) reading test. This was the first year of an assessment program designed to measure school progress toward, among other content areas, three reading outcomes: reading for literary experience, reading to become informed, and reading to perform a task. The MSPAP, given throughout the state, is a nontraditional, criterion- referenced performance assessment, which in 1991 required 9 hours of testing time over an 8-day period. The 3 panels, 1 consisting of experts with an instructional perspective, 1 with a curricular perspective, and 1 a psychometric perspective, independently addressed a variety of test quality issues after review of the test materials and a presentation by a test developer. The results suggest that a test such as the 1991 MSPAP is adequate to assess school progress in reading but may be confounded with writing and may not adequately measure progress in basic reading skills. It was concluded that a test such as the MSPAP is useful for making curricular and instructional decisions, but that use of the test for making decisions about individual students was not supported.*

Several assessment forms that are unlike *traditional*, fixed response (e.g., multiple-choice) tests are becoming increasingly popular in support of school reform efforts (Taylor, 1994), particularly portfolios and performance assessments. Portfolios are collections of examples representative of a student's work over time (Valencia, 1990). Performance assessments involve evaluating students' constructed responses (or demonstrations) in standardized situations (Stiggins, 1988). A variation of performance assessment, which has been called *authentic assessment*, involves tasks that are similar to the processes that are central to a particular discipline (Wiggins, 1989), commonly requiring multiple activities and extended time periods. Arizona, California, Connecticut, Kentucky, and Maryland are examples of states that are engaged in large-scale performance assessment projects (Goldberg & Kapinus, 1993). These forms, together, have been called *alternative* assessments.

There are many advantages of traditional assessments. Primarily, they result from the presence of multiple, independent examinee responses and objectivity of scoring. When examinees respond in closed form, they can be more easily measured under identical conditions. Moreover, using multiple, separate observations, the domain being represented can be sampled efficiently. Reliability is usually high, which means that scores are produced with relatively little error, a crucial characteristic to support "high-stakes" decision making.

Traditional assessments, however, have been criticized. These assessments might not easily, or perhaps even be able to, represent a full range of outcomes that should be expected to result from schooling (e.g., disposition to use prior knowledge in reading, effectiveness of use of prior knowledge in reading, contributing effectively to groups engaged in problem-solving). Because pressure, resulting from high-stakes assessment programs, to increase scores often results in attempts to realign curricula to correspond with the test's domain (Moss, 1994; Shepard, 1990), some professionals believe that, as a result of the assessment, the nature of the reading construct is violated by assuming that discrete skills underlie the reading act. Thus, assessment-driven instruction focusing only on these skills may depend on the untenable assumption that the skills combine spontaneously to produce competent readers. Traditional assessments have therefore been criticized for narrowing the reading curriculum, although anticipated advantages of alternative assessments, for measuring reading or for curricular and instructional practice, have not been established empirically (Hambleton & Murphy, 1992). Also, because a breach of test security would result in bias of the test's domain sampling, the test's items are not commonly available for inspection and therefore cannot be used to help understand the domain that is being measured. As a result, neither teachers nor students may have an exact understanding of what is being measured and therefore may not understand clearly how to improve.

A frequently cited advantage of using alternative assessments centers around the domain that can be measured (Moss, 1994). It is argued that alternative assessments have the capacity to assess complex objectives in a more meaningful and, therefore, more engaging context (Aschbacher, 1991). They can ask students to use higher order cognitive processes, collaboration, and so on, thus measuring a broader domain of activities than traditional assessments. In assessing reading, for example, instruction in relevant prior knowledge may be incorporated prior to engagement in a task. Messick (1994) characterizes arguments favoring the use of performance assessments as based on authenticity (full construct representation, i.e., not leaving any part of the construct out of the assessment) and directness (lack of confounding construct, i.e., not introducing any irrelevant construct into the assessment), arguing that evidence for both is needed.

Other advantages center around the potential open nature of the assessment process. Through alternative assessments, there is an opportunity to describe learning objectives publicly. Thus, criteria (e.g., scoring processes) may be known and understood in advance (Aschbacher, 1991). Using that knowledge,

teachers can effectively modify classroom practices, including incorporating performance assessments modeled after the high-stakes assessments, into their classroom instruction (Baron, 1991). Also, students can know and apply the identical criteria that teachers, districts, and states use, increasing the students' capacity for self-monitoring of learning (Baron, 1991).

Use of large-scale alternative assessments has also been criticized. Primarily, these criticisms center around three areas: low reliability, uncertain validity, and costs. Performance assessments appear to have lower reliability than traditional assessments, particularly across tasks (Dunbar, Koretz, & Hoover, 1991) and occasions, and their accuracy for making high-stakes decisions about individuals has been questioned (Shavelson, Baxter, & Pine, 1992). Their validity is also a concern, because the contexts (tasks, materials, etc.) used in each assessment are restricted and because dependencies of items on these contexts may heavily influence the results (Aschbacher, 1991; Dunbar et al., 1991; Moss, 1994). Moreover, increases in systematic gender and ethnic score differences may be observed (Hambleton & Murphy, 1992), an especially important issue because high-stakes performance assessments will likely be subject to legal scrutiny. Along with bias, other legal areas of concern include contract arrangements, contractor oversight, reliability and validity, opportunity to learn, and cut-score defensibility (Mehrens & Popham, 1992). Regarding costs, some states report that performance assessments can be two to six times the expense of typical standardized tests (Aschbacher, 1991), and that may be an underestimate. Particularly expensive aspects of performance assessment include the need for multiple raters for scoring student responses, training programs for raters in their highly structured scoring systems, validation of complex test content using subject matter-experts, providing students multiple opportunities to pass, equating, and security (Aschbacher, 1991). They also typically require more testing time.

For this paper, the 1991 Maryland School Performance Assessment Program, which included an assessment of reading outcomes, was examined as an example of a large-scale performance assessment in reading. Our purpose was to evaluate this assessment in an attempt to discover what could be useful to other designers of large-scale performance assessment programs in reading. Panels familiar with the 1991 assessment were utilized in an expert review format. Conclusions based on their insights and recommendations are the focus of this study.

THE 1991 MARYLAND SCHOOL PERFORMANCE ASSESSMENT PROGRAM IN READING

In the Spring of 1991, the state of Maryland first administered statewide, in Grades 3, 5, and 8, a criterion-referenced test as part of the Maryland School Performance Assessment Program (MSPAP) that was consistent with many of the alternative assessments being introduced across the country. The test required a total of 9 hours of testing time over an 8-day period and assessed each student's performance in language arts (reading, writing,

and language usage) and mathematics. Other studies of the effects of the 1991 assessment have shown that the MSPAP has affected reading education at the district (Guthrie, Schafer, Afflerbach, & Almasi, 1994) and the individual school (Afflerbach, Guthrie, Schafer, & Almasi, 1994; Almasi, Afflerbach, Guthrie, & Schafer, 1994) levels. This study attempted to evaluate the quality of the 1991 test in reading, grades 3 and 5, for two of its goals: as an assessment of reading achievement, and as a part of a curricular and instructional reform process.

It should be noted that the Maryland State Department of Education (MSDE) has initiated changes in its testing procedures since the 1991 testing. Therefore, this study is restricted to the 1991 test and its administration, and it should not be taken as an evaluation of the MSPAP as it is currently implemented.

These tests are part of the Maryland School Performance Program (MSPP), a data-based system of school measures in areas deemed appropriate for making decisions about school improvement. Eventually, criterion-referenced tests similar to the 1991 assessments will be available in five areas: reading, mathematics, writing and language usage, social studies, and science. Examples of other data-based areas in the system include functional tests in reading, mathematics, writing, and citizenship; promotion and program completion rates; attendance and dropout rates; post-secondary plans and decisions; enrollments; special programs and services; financial data; staffing and instructional time; and norm-referenced test results. Standards are developed for each area that identify levels of satisfactory and excellent performance. Data are aggregated at the school, district, and state levels; reported annually by sex and race/ethnicity; compared with the standards and examined for trends; and used for decision making at all levels of aggregation. Those data-based areas that evaluate student performance were chosen because they are essential and expected of all students, are needed for school improvement, are useful for curricular and instructional improvement, and can be compared with statewide standards (Maryland State Department of Education [MSDE], 1990b).

Goals and Outcomes

The following discussion is intended to provide an overview of the 1991 MSPAP in reading. The test in reading grew out of state-approved learning outcomes, as did all MSPAP assessments. There are four primary goals in reading that apply to all tested grade levels (3, 5, 8, and 11). These are (a) a demonstration of positive attitudes toward reading a variety of texts; (b) a demonstration of ability to construct, extend, and examine meaning for a variety of texts by using strategic behavior and integrating both prior knowledge about reading and topic familiarity; (c) a demonstration of ability to vary orientation by interacting with a variety of texts for different purposes (reading for literary experience, e.g., novels, plays, short stories; reading to be informed, e.g., subject-matter texts, articles, editorials; and reading to perform a task, e.g., follow directions); and (d) a demonstration of ability to interact with a variety of texts and for a variety of purposes through the use of four stances to

construct, examine, and extend meaning. Those four stances are (a) global understanding (considering such things as main theme or topic and author's overall purpose or point of view), (b) developing interpretation (by revisiting the text, clarifying, verifying, and revising understanding by considering such things as plot and character development; by organizing text information; or by following directions to complete a task), (c) personal reflection and response (considering prior knowledge and information from the text through comparing author and self points of view or comparing new and previous background knowledge), and (d) critical stance (identifying and analyzing the author's perspective and craft or the text's mood or clarity) (Goldberg & Kapinus, 1993; MSDE, 1990a).

In scoring the MSPAP, the focus is on three broad outcomes. These are reading for literary experience, reading to become informed, and reading to perform a task (Goldberg & Kapinus, 1993).

Sample Activities

To familiarize teachers with the nature of the MSPAP assessments, MSDE developed samples of activities in the areas tested. Because of test security, the nature of the 1991 test in reading is described here in terms of the sample tasks developed for reading, writing, and language usage in Grade 3 (MSDE, 1991). All questions referred to a story, called "The Quitting Deal," that the students would have read about a mother and daughter making a deal to break their habits (Tobias, 1975). Responses to many of the activity prompts, which were written in a student response booklet, were scored for multiple outcomes. Only scorings for the reading outcomes are described here. It should be noted that the scoring criteria listed were developed as examples of tools but were never used operationally. In practice, scorers were trained to apply elaborations of rubrics similar to these scoring criteria. Most reading items were actually scored using keys with activity-specific descriptors (Goldberg & Kapinus, 1993).

Sample Activity 1 (brief response): Tell a friend in your own words what the story is about. Because your writing will be read by others, be sure that you check carefully for correct spelling, punctuation and capitalization. The scoring criteria for the outcome reading for literary experience—global understanding stance were *0* for no response, *1* for an attempt not to the point, *2* for a partial response that mentioned one of two major story elements but not both, or *3* for a complete response that mentioned both major elements.

Sample Activity 2 (brief response): Why are the characters in the story trying to break their habits? The scoring criteria for the outcome reading for literary experience—developing interpretation were *0* for no response, *1* for poor but still correct (e.g., they are bad), *2* for satisfactory (e.g., their habits are bad for them), or *3* for a response that described why each character is trying to break her habit.

Sample Activity 3 (brief response): Pretend that you and your friend both want to quit a habit. Think about a deal that you and your friend might make. Do you think your deal would work better than the deal Jennifer and her mother made? Why or why not? Write a

few sentences, being sure to use information from the story to explain your answer. The scoring criteria for reading for literary experience—personal response stance were *0* for no response, *1* for an incorrect attempt that did not relate information from the story to the student's own deal, *2* for a student's own deal that was generally related to the story, or *3* for a student's own deal that was explicitly related to information from the story.

Sample activity 4 (brief response): What in the story tells you if the author thinks it's better to solve problems with the help of another person or by yourself? The scoring criteria for reading for literary experience—critical stance were *0* for no response, *1* for an attempt that was incorrect or unrelated to the story (e.g., it's better to have someone's help), *2* for a response that was related to the story in general (e.g., it's better because the girl and her mother help each other), or *3* for a response that was related to specifics in the story.

Sample activity 5 (extended response): The story tells how Jennifer and her mother tried to break their habits. Use your imagination and information from "The Quitting Deal" to write a story or a poem that tells about a cure for Jennifer or her mother that is different from the ones that they tried in the story. If you wish, you can use a picture to illustrate your writing in the space provided on the last page of the student response book. The scoring criteria for personal response stance were as follows. A score of *0* for no response or one that could not be read, or one that did not address the question or was unrelated to the task. A score of *1* for a response that showed little understanding of the text, did not include relevant text features as supporting evidence or used examples unrelated to the topic, was superficial or overly general, indicated little or inaccurate inferencing, copied directly from the text, and made no attempt to synthesize information or ideas within the elements of the text or across texts. A score of *2* for a response that relevantly but inconsistently related personal experience and/or prior knowledge to the text, included some examples from personal experience and/or knowledge and from the text using relevant text features, revealed literal understanding but little or no evidence of abstraction, and attempted to provide links between personal experience and the text, but in which the links were not always clear, consistent, and coherent. Lastly, a score of *3* for a response that consistently and relevantly related personal experience and/or prior knowledge to the text; included extensive examples from personal experience and/or knowledge and from the text utilizing relevant text features; contained evidence of abstraction; and provided clear, consistent, and coherent links between personal experience and the text.

METHOD

This study utilized three expert panels to review the 1991 MSPAP reading assessment. The panels met on May 8, 1993. The materials, participants, and activities are described in the following sections.

Materials

Prior to the day of the review, each panel member received an overview of the framework for reading outcomes that was the basis for the test, including goals and objectives, a

secure copy of the test along with a description of how the scoring was accomplished, and a list of the questions that the panels were asked to answer on the date of the review. The questions, from a test-review questionnaire of our own design, were given in the form of statements about which each panel was asked to agree or disagree (or indicate insufficient information) and to provide comments. In addition, the two members of the psychometric panel (see following section) were provided copies of the technical report for the test (CTB Macmillan/McGraw-Hill, 1992).

Participants

Each panel was composed of persons who were familiar with the 1991 testing in Maryland. Three members of our research team acted as facilitators for the groups and recorded the responses for later analysis.

One panel included a district elementary supervisor, a district coordinating supervisor, a school-based reading specialist, and a district reading specialist. This panel was constituted primarily to provide insights from the perspective of practitioners. It is identified here as the *instructional panel*.

A second panel included two university faculty members in human development, and an elementary school principal. This panel was constituted primarily to provide insights from the perspective of educational psychologists who work in the area of reading development. It is identified here as the *curriculum panel*.

The third panel included two university faculty members in measurement, statistics, and evaluation. This panel was constituted primarily to provide insights from the perspective of psychometric specialists and is identified here as the *psychometric panel*.

Other participants were a reading specialist instrumental in the development of the 1991 reading assessment at MSDE, who was available to the panels for clarification about the development and implementation of the 1991 assessment, and a psychometrician at MSDE who provided technical support for the psychometric panel. A fourth member of our research team was available to respond to procedural questions and otherwise to coordinate the panels' work.

Activities

The three panels met for a 1-day test review. Following an introduction to the task and a brief overview of the tests, the panels met separately to complete the reviews. Because of time constraints, it was anticipated that not all panels would be able to respond to all questions in the test-review questionnaire. Therefore, each panel was asked to make sure it responded to items that had been identified previously as within its expertise and was invited to comment on other areas as desired and able. The panels were asked to reach a consensus about each item. The review team member noted the points raised during discussion by each panel for later analysis. The panels worked independently, and each panel was able to address its questions and several others during the 1-day review period.

RESULTS

The findings of the panels are described together according to the statements in the test-

review report form. In response to each statement, the panels were asked to express agreement or disagreement (or to indicate that there was insufficient information) and were invited to provide additional comments. In the results that follow, the panel identifications indicate which of them responded to that statement. The statements, themselves, are grouped into areas of commonality.

Area 1: Content Domain (Behaviors) and Sampling

Statement 1A: The domain of content and skills the test is intended to measure has clearly been defined. The three panels agreed with this statement. Although believing that the specifications of content and skills were thorough and well developed and were useful to teachers and supervisors, several concerns were expressed. Some thought the taxonomy of skills seemed like a "laundry list," and that the boundaries of the taxonomy were sometimes fuzzy (e.g., what is the difference between the global understanding and the developing interpretation stances). Concerns were expressed about the stances lack of narrative description and about some combinations of reading purposes and stances (e.g., developing global understanding or critical understanding while reading to perform a task is an unusual notion). Questions were raised about whether teachers had learned about the outcomes and stances presented in the reading model.

Some panelists expressed concern about the emphasis on productive aspects in evaluation of reading as opposed to recall and recognition. The panels thought that achievement was underestimated because students were evaluated only on what they wrote or otherwise committed to paper (such as drawings or diagrams). Students' ability to read words, such as whether they have attained requisite decoding and phonics skills, was not assessed. Thus, the whole language philosophy of the outcome model may depress students' and schools' scores.

The panels thought the act of answering a question may prompt deeper understanding. Thus, one cannot necessarily make the inference that the students actually did the things they were being scored on at the time that they were reading the passage.

Statement 1B: The description is complete enough to allow determination of whether the content of the test matches a given curriculum or area of study. The panels were mixed: Two expressing agreement and the curriculum panel disagreement. The curriculum panel believed the description was too inferential; that without a narrative elaboration of the model, it was difficult to know whether a specific curriculum matched the stances.

Statement 1C: The emphasis given to each content and skill area in the test has been described clearly. The three panels disagreed with this statement. They believed the emphasis given to each stance would not be clear, particularly to teachers and to parents.

Area 2: Task Domain

Statement 2A: There is a rationale for the tasks included in the test. The three panels disagreed with this statement. Although they presumed the test developers had a rationale, they

did not have it to review. They suggested that a rationale be included in the test administration manual, describing for each prompt what reading purposes and stances are being assessed.

Statement 2B: The description allows determination of whether the tasks match a given curriculum or area of study. Two panels agreed with this statement. The curriculum panel disagreed and found no discussion of curriculum; some believed that teachers did not have easy access to curricular documentation that may have existed. It was suggested that the tasks be made available to teachers as a form of feedback to the extent allowed by test security, or at least that more extensive prototypes, including sample tasks and scoring guides, be developed and made available.

Area 3: Adequacy of Task Sampling

Statement 3A: The test is a representative sample of the specified domain. The psychometric panel expressed agreement, the curriculum panel disagreement, and the instructional panel indicated insufficient information in response to this statement. The instructional panel thought that not all stances were covered for all text types, but that other test forms might have included the missing domain elements. The curriculum panel made similar comments, but also thought that some skills, such as decoding, were not assessed. They thought they would have to decide which area each item assessed and then count them up across the test forms to obtain a judgment of breadth, but they did not take the time to do that. It should be mentioned, however, that the test forms were constructed to be roughly parallel in 1991; since that time, they have become complementary, to sample a broader domain. The curriculum panel also found no evidence that the stories represented the content domain. The psychometric panel based their agreement on the ability of matrix sampling to permit a broad spectrum of tasks to be used in the assessment of a school. This was done in the 1991 MSPAP only for writing, however.

Area 4: Clarity of Tasks

Statement 4A: The tasks are clearly written for the intended age level. The instructional and curriculum panels disagreed with this statement. They believed that one had to assume the stories were selected to be appropriate for the age levels of the tests. Moreover, different types of print sizes and illustrations may have affected the results.

Students who became visibly upset during the test administration were excused. The instructional panelists believed, however, that it was unrealistic to expect third-grade students to read two texts and to do 13 complex and demanding independent tasks, working without teacher help for 50 minutes. Any third-grader who could not read was required to sit through the entire session, amplifying his or her frustration and sense of failure. Noting the target year of 2000 for satisfying state standards, one panelist observed that "in 2000, 9-year olds will still be 9-year olds."

Statement 4B: The tasks are consistent with the objectives of the test. The curriculum panel agreed with this statement. The instructional

panel indicated insufficient information, believing the objectives ambiguous, and that the answer might be different depending on whether the objectives referred to the assessment of the reading domain or to the fundamental objectives of the MSPAP (i.e., school improvement).

Area 5: Clarity of Directions for Administration

Statement 5A: The directions for administration are clearly written. The three panels agreed with this statement. The instructional panel believed, however, there was too much emphasis on test security. Although the MSPAP process has changed since 1991, for that year, there was no teacher preview of the test or of the administration manuals and materials until the actual day of testing. The panelists wondered what message was being sent to the students about the value and importance of the test if the teacher's attitude was negative.

Statement 5B. The test can be easily and effectively administered by teachers in the prescribed manner. The curriculum panel agreed with this statement, and the instructional panel disagreed. The instructional panel cited lack of familiarity with the test beforehand on the part of the teachers and lack of clarity about whether directions may be repeated. They thought there were too many directions for the students to understand and work on for long periods of time without follow-up. They cited the length of the testing time as a problem, because by the end of the period, the students were probably exhausted by the process and may not have performed well even if they were capable. They also thought the test was not sufficiently resistant to an unexpected disruption that may occur during administration.

Statement 5C: The administration instructions are sufficiently standardized, so that different administrators may be expected to elicit comparable responses. All panels agreed with this statement.

Area 6: Test Development

Statement 6A: Adequate procedures were used to develop the tasks. The three panels agreed with this statement. The psychometric panel was particularly positive about the involvement of a wide range of teachers in the development process.

Statement 6B: Appropriate procedures were used to try out (pilot) the tasks. The psychometric panel disagreed and the curriculum panel agreed with this statement. Nonsystematic piloting of tasks was carried out in Delaware and Philadelphia, but the psychometric panel expressed concern that no formal statistical analysis of the pilot data was done. The panel believed that, had analysis been done, excessive difficulty in the core booklets might have been identified. The psychometric panelists also were concerned about low motivation to perform well on the part of students who participated in the pilot data collection.

Statement 6C: Appropriate procedures were used to try out the directions for administration. The psychometric panel agreed with this statement but felt the pretesting of the directions was somewhat limited in its ability to provide adequate information about consistency across administrators.

Statement 6D: Results from tryouts were used to improve the test. The psychometric

panel disagreed, believing that the improvements made were limited by the lack of sufficient data from the pretesting.

Area 7: Statistical Characteristics

Statement 7A: Appropriate procedures were used to evaluate score reliability. The psychometric panel had available a technical report carried out by CTB Macmillan/McGraw-Hill (1992). Nevertheless, they thought there was insufficient information to evaluate this statement. They thought the level of the reported reliabilities was poorly documented (i.e., it was not clear whether the reliabilities reported, alpha homogeneity coefficients, were for students, for classrooms, or for schools).

Statement 7B: Score reliability is adequate for the purpose of the test. The psychometric panel agreed, indicating that the coefficients of rater consistency indicated a high degree of objectivity in the scoring, and that the mean scores for schools were adequately generalizable across raters.

Statement 7C: Appropriate procedures were used to provide statistical evidence of validity. The psychometric panel agreed but thought the procedures used to evaluate validity were weak. They noted that the report did not provide statistical evidence to support a judgment about content validity of the items as they contribute to the construct validity of the scores. Although the technical report contained frequent references to a method to evaluate dimensionality, the statistic was not reported, and no other evidence was presented to evaluate the appropriate number of factors. Thus, the panel believed there was no empirical motivation for examining a two-factor solution, and yet only a two-factor solution was presented ("Why not one factor or three?").

Statement 7D: The statistical evidence of validity is adequate for the purpose of the test. The psychometric panel disagreed, because they thought the factor analysis did not clearly support differentiation of the constructs. The panel suggested that further work on clarification of the constructs underlying the test and providing statistical evidence of their justification was needed. They also thought, however, given that the test design was new, that this sort of finding was not surprising and should not be taken as a negative judgment.

Statement 7E: Adequate procedures were used to evaluate subscore relationship. The psychometric panel agreed and thought the multitrait-multimethod analysis done was adequate for this purpose.

Statement 7F: Subscores are adequately independent for the test's purposes. The psychometric panel disagreed. Along with an overall reading score, subscores were reported for three scales: reading as a literary experience, reading for information, and reading to perform a task. Although the subscores were measured independently, the panel thought there was no empirical evidence to support the distinction. They observed that the correlations between reading and writing, which were very near their reliabilities, suggested little distinction between these constructs. They also thought that there seemed to be much more dependency between language and mathematics scores on the MSPAP data than, for purposes of comparison, on the California Test of Basic Skills.

Statement 7G: Adequate procedures were used to evaluate equivalency of forms. The psy-

chometric panel agreed but thought the method of sampling for forms was problematic, because only larger schools received extra forms.

Statement 7H: The forms are sufficiently equivalent for the purposes of the tests. A generalizability study was performed across schools. The psychometric panel agreed with this statement, based on the results of this analysis and the interrater consistency data.

Statement 7I: Most examinees have sufficient time to complete the test. The psychometric panel expressed insufficient information, because there were no statistical data presented to evaluate this statement. The instructional panel disagreed with the statement, based on logical analysis. They thought the time frame was unrealistic and wondered why it should be a timed test at all. Moreover, they thought the demand that writing revisions be done during the same time frame that the original writing took place is not reflective of the way in which writing is typically done.

Area 8: Fairness

Statement 8A: There are no irrelevant sources of difficulty (or easiness) that would seriously affect the test scores. The curriculum panel disagreed with this statement. They thought the difficulty of the questions and their wording raised cultural fairness concerns. They also believed that if a student who missed a day of the test and then had to go back and read something, he or she did not have the same level of understanding as one who had written about the issues or discussed them. They were concerned that there were no opportunities for make-up testing. They thought that writing and drawing were irrelevant sources of difficulty in the assessment of reading, as were double-scoring of responses and the ability to follow instructions. They were also concerned about the ability of students to manipulate the materials and to read and follow instructions for several tasks at once. It was suggested that scores be reported separately for English as a secondary language (ESL) and non-ESL students.

The mixing of students from different classrooms for testing was seen as undesirable for two reasons. First, some students may have lost confidence when taken out of their regular classes. Second, students may have been grouped with nonoptimal peers for activities that include small-group interactions.

Statement 8B: The test contains no potential sources of bias against specific groups. The curriculum panel agreed with this statement. They noted, however, that drawing analogies, making inferences, and comprehension is easier when the information is familiar. In an effort to avoid bias, the passages were neutral to ethnic groups and became somewhat generic. This may have made it difficult to apply relevant background knowledge. It is also not clear that all ethnic groups learn or take tests best in the same way.

Statement 8C: The test contains no material that particular groups of persons may find offensive. The curriculum panel agreed.

Area 9: Scoring

Statement 9A: The scoring tools (keys, rules, and rubrics) are appropriate for the purposes and tasks of the test. The instructional panel agreed with this statement. They felt some items, however, did not adequately

prompt the students to write their responses with effective knowledge of the criteria by which they were eventually scored. They also were concerned that the categories of stance were not mutually exclusive, and that the tasks may have included more than one stance but had been scored using rubrics that were exclusively for one stance (it should be noted that most of the reading items were scored with keys that were not stance specific). This could have resulted in a task-rubric mismatch and may have introduced "noise" in the scoring process. It was possible, though, that the scorer training process could have addressed this concern.

Statement 9B: The procedures for training of raters are adequate. The instructional panel felt insufficient information was available to evaluate the training procedures. The description of the training process was not adequate to determine what actually happened. They thought the interrater correlations suggested the training was adequate, however.

Statement 9C: The procedures for selection and screening of raters are adequate. The instructional panel disagreed with this statement. They did not feel the invitation to participate was systematic or that a wide range of means of inviting people to be selected was used. They noted that different means of communication were used in different districts and in different schools. The panel suggested that information about nomination and selection of raters be included routinely in the test information materials for teachers who administer the test.

Statement 9D: The procedures for assuring comparability of ratings are adequate. The instructional panel also disagreed with this statement. Ratings were done in California as well as Maryland. Given that only half of the raters in Maryland achieved the MSPAP standard of 70% exact agreement of scores, they questioned whether the procedures were adequate. They also wondered whether the California raters differed in ways other than increased level of agreement across raters. Although the scorings in the two states were equated, that was necessitated because consistent differences were observed in overall ratings, perhaps because of different scorer training procedures.

Statement 9E: The scoring procedures are sufficiently standardized so that different scorers may be expected to arrive at comparable scores. Although believing that the scoring procedures looked good on paper, on the basis of low agreement achieved between raters, the instructional panel disagreed that the scoring procedures were sufficiently standardized. They also wondered whether raters could separate reading and writing in scoring.

Statement 9F: There is adequate evidence that different scorers may be expected to arrive at comparable scores. The instructional panel disagreed with this statement, because only 50% of the readers in Maryland had met the minimum standard of 70% exact agreement.

Area 10: Norms

Statement 10A: It would be meaningful to have normative data for this test. The curriculum panel agreed with this statement. They felt it would provide a comparison with other groups useful to a school planning change.

Statement 10B: If norms would be meaningful, normative data are provided (may have

resulted from an actual administration). Normative data were not provided. The curriculum panel felt that the time delay in getting data from the MSPAP administration was too long, and that the data were not reported in a way that is helpful in instructional planning. Interpretations were made relative to proportions of students in various levels of performance, but the effectiveness of the information depended on where the cut scores between the performance levels were.

Area 11: Security

Statement 11A: The provisions for test security are adequate. The curriculum panel agreed with this statement but believed the test was secure to an undesirable extent. It was impossible to retest, so students who were absent received scores of zero. The panel felt that restriction of access to the test meant teachers could not evaluate what was and was not covered. They wondered whether the designers of the test were the state's best reading experts, noting that people knowledgeable about education and educational processes were not included. They also wondered if prior knowledge about the test might have been helpful to those teachers who participated in its development; but if what was tested was to be obvious from the published descriptions, they wondered why the test should be secure.

Statement 11B: The procedures for ensuring security of test results are adequate. The curriculum panel agreed with this statement.

Area 12: Standards

Statement 12A: The procedures used to establish criterion scores for test interpretation are adequately described. In setting standards, criterion scores were based on a process of data-guided judgments. A technical description, evaluated by the psychometric panel to be adequate, follows.

In reading, 15% of the items were dichotomies; these were scaled using the three-parameter logistic model (Lord, 1980). Another 66% were three-category and another 15% were four-category, along with another 3% that consisted of items judged to be dependent and therefore grouped together for scaling; these were scaled using a two-parameter partial-credit model (Bock, 1972). Using these calibrations, both item category locations and examinees were placed on a logit scale. Transformation of the logit scale resulted in a scale score with a mean of about 500 and a standard deviation of about 50. An examination of the information values of the items suggested that about one-third of them could be deleted, and following verification that no content gaps would result, they were dropped from the standard-setting process. Five proficiency levels (5 = low, 1 = high) were then established in terms of scale scores. Level 4, "minimal," was established at about 490 and Level 1, "highly advanced," at about 620, based on the availability of interpretable item score information and consistency with the other content areas assessed. Level 5 was below 490. Intermediate proficiency levels were also established: Level 3, "basic," or "proficient," at a score level of about 530 and Level 2, "advanced," at about 580. Although the majority of the examinees fell in levels 4 and 5, these outcomes of the standard setting process were judged acceptable to set high achievement expectations for the outcomes measured. The

proficiency levels were then described by a committee of content experts based on scoring anchors for item score categories that typified proficiency-level locations on the score scale. For example, in reading, "regardless of grade level, highly advanced readers '. . . construct, extend and examine the meaning of grade appropriate texts by making judgments, connections, and extensions of the text that are substantially supported.' Minimal readers '. . . make limited, relevant inferences with implied text support.' The key to the grade-to-grade differences in Reading lies in the term 'grade appropriate texts'" (CTB Macmillan/McGraw-Hill, 1992, p. 11-3).

Statement 12B: The procedures for establishing criterion scores for interpretation are adequate for the stated purposes of the test. The psychometric panel agreed with this statement. They noted that the cut scores were developed empirically and that content characterizations of them were tailored to describe the score levels. They believed this was an excellent approach to integrating norm-referenced and criterion-referenced assessment.

Statement 12C: There is adequate evidence to support the use of the criterion scores for interpretation used for this test. The psychometric panel agreed with this statement.

Area 13: Utility

Statement 13A: The uses to be made of the test results are adequately described. The psychometric panel agreed and the curriculum panel disagreed with this statement. The curriculum panel thought the primary use was accountability. They thought the score reports too slow, however, and not amenable to principals working with staff for improvement; more specific information is necessary. Use of the data at the student level, which was not encouraged by MSDE, was specifically discouraged by the psychometric panel because of nonequivalence of forms and low student-level reliability.

Statement 13B: The test quality is sufficient to support those uses. The curriculum and psychometric panels thought the test quality was adequate, but the curriculum panel believed the reporting of the results needed to be more timely. This problem stemmed from the labor-intensive nature of the scoring, which arose because of the nature of the tasks. Although the curriculum panel thought that information about individual student and teacher performance would be helpful in instructional decision making, they agreed with MSDE that such information might easily lead to misuse of the test results.

Statement 13C: All responsible uses of the test results have been anticipated and described adequately. The psychometric panel agreed and the curriculum panel disagreed with this statement. Although the curriculum panel thought many uses were covered, they thought only time would tell if other uses emerge.

Statement 13D: There is adequate protection against potential misuses of the test results. The curriculum panel disagreed, noting that misuses were not mentioned in the written materials. They were concerned that the test may not be of adequate quality to justify a decision by the state to take over a school, and that therefore this may constitute a misuse. They also wondered whether there was adequate protection against making judgments about individual teachers or students on the

basis of this test, particularly because no other test was mandated. They were concerned about whether comparisons of schools with dissimilar characteristics were appropriate. They expressed distrust that the motivation for the MSPAP was political and prompted by a lack of confidence in teachers and a concern for placing blame. The panels observed that teachers would, indeed, be blamed for things that were not their problems. Noting that an assumption of MSDE in its school reform process was that "every child can learn," one panelist observed that "with a mind set that 'every child can learn,' we will not find reasons why some children cannot."

Statement 13E: The score reporting method is appropriate for the purposes of the test. Scale scores for reading were reported for each student on electronic media. For each student, based on his or her position on the logit scale, an expected percent of maximum (EPM) score was also estimated and reported for each reading outcome across all test forms using the item parameter estimates. Use of this information for decision making about individual students was discouraged.

Hard-copy reporting was aggregated at the school, system, and state levels. The report included percentages of students at each proficiency level and, for each outcome, the mean, median, and standard deviation of EPM scores and the percentage of students scoring in each EPM quartile.

The curriculum panel disagreed with this statement. They thought the information was too late and not precise enough to be useful in guiding school improvement. They raised concern about using disaggregated data by various types of students. They also thought that the scores of transfer students should be treated separately in school-level reporting.

Statement 13F: The score reporting method is clear to the intended audience. The curriculum panel disagreed with this statement. Using schools as the intended audience, they cited deficiencies in the areas of lateness of reporting and unclear definitions of the five levels.

Statement 13G: The score reporting method is useful for educational decision makers. The curriculum panel also disagreed with this statement. They noted a lack of information about how to improve scores and about scores in the specific areas of the reading domain of the test.

Area 14: Judgment

Statement 14A: This test is adequate for its stated purposes. The instructional and psychometric panels judged the test adequate, and the curriculum panel did not. The psychometric panel noted that this was a very responsible execution and analysis of a set of authentic reading assessment tasks, and that the MSPAP appears to be well on target.

The instructional panel thought that if the tasks represented the domain and set of behaviors, then it was adequate, except that reading to perform a task should have been represented. It was suggested that perform-a-task scenarios be developed and related to reading. They noted, however, that all reading accomplishment was assessed through writing, drawing, or graphic organizers. If the rationale for this is not available, it is difficult to assess the adequacy of the test for its purposes.

The curriculum panel cited deficiencies in two areas. First, they thought the test was

likely to underestimate performance because of its length, complexity, confounding of writing with reading, and lack of assessment of fundamental skills such as decoding. Second, they thought the reporting methods were inadequate to guide instruction.

CONCLUSIONS

The points made in the panels' reviews focused on the MSPAP and particularly the 1991 administration. Many, therefore, do not apply to other, similar efforts or even to the current MSPAP. There do appear to be several conclusions, however, that may be drawn from the panels' judgments.

None of the panels thought the assessment was either trivial or off-target, with the exception that basic skills are untested. The implementation of a test such as the 1991 MSPAP appears to have the potential to provide motivation for reforms in reading curriculum and instruction. Other studies that have described changes as a result of the 1991 MSPAP at the district (Guthrie et al., 1994) and at the school and classroom levels (Afflerbach et al., 1994; Almasi et al., 1994) are consistent with this conclusion.

The domain of the MSPAP, as described in the reading outcomes model, was evaluated by the panels as clearly defined. Whereas the panels suggested some specific ways in which the model could be improved, it nevertheless appears adequate to provide direction to reform efforts in reading, guided by accountability assessment through broadly identifying goals for change. As Taylor (1994) has noted, many professional groups are working on domain descriptions in various content areas. Elaborations of these domains in terms of assessment contexts (e.g., tasks) that are consistent with current understandings in cognition, learning, motivation, and instruction (Baron, 1991) and scoring systems (e.g., scales, deys, rubrics, protocols) that can be understood and applied by teachers, students (Baron, 1991), and parents would help to operationalize the domain specifications.

The panels found both logical and statistical evidence of confounding of reading results with other outcomes, particularly writing, in the assessment of the domain. Subscale independence can be particularly troublesome in a test such as the 1991 MSPAP that uses writing as the sole means of obtaining evidence about reading performance, and particularly for assessments that feature multiple scorings of responses for different educational outcomes.

A test such as the 1991 MSPAP appears to have adequate ability to assess school progress in reading. The psychometric panel felt that the statistical analysis of the test supports making judgments on the basis of scores at the school level. It should be noted, however, that several panelists thought the lack of interpretable student-level scores was a severe deficiency in the 1991 MSPAP. On the basis of the psychometric analysis, we agree that interpretations should not be attempted at the student level, and that uses of the scores resulting from the assessment should be restricted to making school-level judgments. Perhaps more reliable and more valid student-level data could be generated by including a broader range of responding and scoring mechanisms in large-scale testing programs such as the MSPAP.

In summary, we conclude that a program such as the 1991 MSPAP can be useful for

both curricular and instructional decision making. Curricular decisions may be made based on a domain description such as the MSPAP reading outcomes model. Instructional decisions may be based on the texts, tasks, and responses expected from the student. Detailed reporting and supporting material would be necessary to support extended systemic change, however.

Suggestions for Test Developers

The panels were asked to provide recommendations for developers of large-scale performance assessments similar to the MSPAP. We end with a summary of the suggestions of all three panels divided into activities before, during, and after testing, as recommended by one of the panels. It is suggested that others planning a similar assessment activity give thought to these points that were developed by persons who have a variety of perspectives from which to react to their first-hand experiences with the MSPAP.

Prior to test administration

Allow teachers to preview the test prior to administration.

Familiarize teachers with the administration procedures.

Clarify the test's purposes, audience, uses, and misuses.

Make the rationale for the test explicit, including its intended uses.

Evaluate the test for whether it is developmentally appropriate at each grade level.

Minimize confounding of writing and reading (e.g., add multiple-choice items or oral responding).

Find a way to assess that will be useful for decisions about individuals (e.g., placement).

Describe the curriculum that will be tested to teachers and administrators.

Involve business leaders in the development of the curriculum and the assessment.

Pilot the test (items and administration procedures) extensively and analyze the results fully.

Study the effects of the various test factors to evaluate generalizability.

Provide test and test form specifications in writing.

Survey curriculum-assessment overlap formally.

During test administration

Use grade-appropriate numbers of tasks (particularly important for lower grades).

Use matrix sampling of content to cover the domain.

Include teacher direction of activities.

Avoid making students who cannot perform just sit for long testing segments.

Provide a way to include students who are absent in the assessment.

Test early in the year (e.g., May can be too hot and is a time when schools are readying to close).

Eliminate the effects of time constraints.

Allow teacher-student interaction.

Keep students in their regular classrooms.

Test a sample of students in each school (assuming only school-level data is used).

Test groups of schools on a schedule over some number of years.

Locate passages, tasks, and items so that they are readily found by the students.

Provide a convincing motivation for students to do well on the test.

After test administration

Report all information that has appropriate uses.

Report the information soon, so it is timely and meaningful.

Include individual information in the score reports if it is interpretable.

Explore a variety of useful formats for the information (so educators will understand them).

Clarify psychometric test reports to distinguish between student and school-level statistics.

Author Note. The authors are grateful to Steven F. Ferrara and Gail Lynn Goldberg for helpful comments on an earlier draft.

REFERENCES

Afflerbach, P. P., Guthrie, J. T., Schafer, W. D., & Almasi, J. F. (1994, April). *Barriers to the implementation of a statewide performance assessment program.* Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Almasi, J. F., Afflerbach, P. P., Guthrie, J. T., & Schafer, W. D. (1994, April). *Impacts of a statewide performance assessment program on classroom instructional practice.* Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Aschbacher, P. R. (1991). Performance assessment: State activity, interest, and concerns. *Applied Measurement in Education, 4*(4), 275–288.

Baron, J. B. (1991). Strategies for the development of effective performance exercises. *Applied Measurement in Education, 4*(4), 305–318.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29–51.

CTB Macmillan/McGraw-Hill. (1992). *Final technical report Maryland school performance*

assessment program 1991. Unpublished document.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, *4*(4), 289–303.

Goldberg, G. L., & Kapinus, B. (1993). Problematic responses to reading performance assessment tasks: Sources and implications. *Applied Measurement in Education*, *6*(1), 281–305.

Guthrie, J. T., Schafer, W. D., Afflerbach, P. P., & Almasi, J. F. (1994, April). *District-level policies of reading instruction in Maryland and their relation to the statewide performance assessment*. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Hambleton, R. K., & Murphy, E. (1992). A psychometric perspective on authentic measurement. *Applied Measurement in Education*, *5*(1), 1–16.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Maryland State Department of Education. (1990a). *Learning outcomes in mathematics, reading, writing/language usage, social studies, and science for Maryland school performance assessment program*. Unpublished document.

Maryland State Department of Education. (1990b). *Maryland school performance program state data-based areas* (Publication 7.5). Baltimore, MD: MSDE.

Maryland State Department of Education. (1991). *Sample activities, student responses, and Maryland teachers' comments on a sample task reading/writing/language usage grade 3*. Unpublished document.

Mehrens, W. A., & Popham, W. J. (1992). How to evaluate the legal defensibility of high-stakes tests. *Applied Measurement in Education*, *5*(3), 265–283.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13–23.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, *23*(2), 5–12.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher*, *21*(4), 22–27.

Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching the test? *Educational Measurement: Issues and Practice*, *9*, 5–14.

Stiggins, R. J. (1988). The design and development of performance assessments. *Educational Measurement: Issues and Practice*, *6*(3), 33–42.

Taylor, C. (1994). Assessment for measurement or standards: The peril and promise of large-scale assessment reform. *American Educational Research Journal*, *31*, 231–262.

Tobias, T. (1975). *The quitting deal*. New York: Viking Press.

Valencia, S. W. (1990, January). A portfolio approach to classroom reading assessment: The whys, whats, and hows. *The Reading Teacher*, *43*, 338–340.

Wiggins, G. (1989, April). Teaching to the (authentic) test. *Educational Leadership*, *46*, 41–47.

[redacted]

[redacted]

NRRC National
Reading Research
Center

318 Aderhold, University of Georgia, Athens, Georgia 30602-7125
3216 J. M. Patterson Building, University of Maryland, College Park, MD 20742